DOCUMENT RESUME

ED 303 518 TM 012 785

AUTHOR Federico, Pat-Anthony; Liggett, Nina L.
TITLE Computer-Based and Paper-Based Measurement of Semantic Knowledge.
INSTITUTION Navy Personnel Research and Development Center, San Diego, Calif.
REPORT NO NPRDC-TR-89-4
PUB DATE Jan 89
NOTE 34p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Computer Assisted Testing; *Knowledge Level; Military Personnel; *Semantics; Test Construction; Test Format; Test Reliability; *Test Validity
IDENTIFIERS Internal Consistency; Navy; *Paper and Pencil Tests; Threat; Weapons

ABSTRACT

Seventy-five subjects (Naval F-14 and E-2C crew members) were administered computer-based and paper-based tests of threat-parameter knowledge represented as a semantic network in order to determine the relative reliabilities and validities of these two assessment modes. Estimates of internal consistencies, equivalences, and discriminant validities were computed. It was established that: (1) computer-based and paper-based measures (i.e., test score and average degree of confidence) are not significantly different in reliability or internal consistency; (2) for computer-based and paper-based measures, the average degree of confidence has a higher reliability than does the average response latency, which in turn has a higher reliability than do the test scores; (3) a few of the findings are ambivalent since some results suggest that equivalence estimates for computer-based and paper-based measures (i.e., test score and average degree of confidence) are about the same, and another suggests that these estimates are different; and (4) the discriminant validity of the computer-based measures was superior to that of the paper-based measures. The results support the findings of only some of the research. This highlights the fact that the reported literature on this subject is contradictory and inconclusive. Seven tables present study data. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Navy Personnel Research and Development Center

San Diego, CA 92152-6800 TR 89-4 January 1989

ED303518

Computer-Based and Paper-Based Measurement of Semantic Knowledge

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P-A. FEDERICO

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Approved for public release; distribution is unlimited.

TM012785

BEST COPY AVAILABLE

TR 89-4 January 1989

COMPUTER-BASED AND PAPER-BASED MEASUREMENT OF SEMANTIC KNOWLEDGE

Pat-Anthony Federico
Navy Personnel Research and Development Center

Nina L. Liggett
University of California, San Diego

Reviewed and approved by
E. G. Aiken

Released by
B. E. Bacon
Captain, U. S. Navy
Commanding Officer
and
J. S. McMichael
Technical Director

Approved for public release;
distribution is umlimited.

Navy Personnel Research and Development Center
San Diego, California 92152-6800

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NPRDC TR 89-4	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Navy Personnel Research and Development Center 6b OFFICE SYMBOL (If applicable) Code 15	7a NAME OF MONITORING ORGANIZATION
6c ADDRESS (City, State, and ZIP Code) San Diego, CA 92152-6800	7b ADDRESS (City, State, and ZIP Code)
8a NAME OF FUNDING / SPONSORING ORGANIZATION Office of Naval Technology 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (City, State, and ZIP Code) Washington, DC 20350-2000	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 62233N PROJECT NO: RF62-522 TASK NO: 801-013 WORK UNIT ACCESSION NO: 03.04

11 TITLE (Include Security Classification)
Computer-Based and Paper-Based Measurement of Semantic Knowledge

12 PERSONAL AUTHOR(S)
Federico, Pat-Anthony and Liggett, Nina L.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Technical Report	FROM 86 Oct TO 88 Mar	1989 January	33

16 SUPPLEMENTARY NOTATION

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	09	

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Computer-based testing, measurement, assessment, modes of assessment, test-item administration

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Seventy-five subjects were administered computer-based and paper-based tests of threat-parameter knowledge represented as a semantic network in order to determine the relative reliabilities and validities of these two assessment modes. Estimates of internal consistencies, equivalences, and discriminant validities were computed. It was established that (a) computer-based and paper-based measures, i.e., test score and average degree of confidence, are not significantly different in reliability or internal consistency; (b) for computer-based and paper-based measures, average degree of confidence has a higher reliability than average response latency which in turn has a higher reliability than the test score; (c) a few of the findings are ambivalent since some results suggest equivalence estimates for computer-based and paper-based measures, i.e., test score and average degree of confidence, are about the same, and another suggests these estimates are different; and (d) the discriminant validity of the computer-based measures was superior to paper-based measures. The results of this research supported the findings of some studies, but not others. As discussed, the reported literature on this subject is contradictory and inconclusive.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION	
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED	
22a NAME OF RESPONSIBLE INDIVIDUAL Pat-Anthony Federico	22b TELEPHONE (Include Area Code) (619) 553-7777	22c OFFICE SYMBOL Code 15

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

FOREWORD

This research was performed under Exploratory Development work unit RF63-522-801-013-03.04, Testing Strategies for Operational Computer-based Training, under the sponsorship of the Office of Naval Technology, and Advanced Development project Z1772-ET008, Computer-Based Performance Testing, under the sponsorship of Deputy Chief of Naval Operations (Manpower, Personnel, and Training). The general goal of this development is to create and evaluate computer-based representations of operationally oriented tasks to determine if they result in better assessment of student performance than more customary measurement methods.

The results of this study are primarily intended for the Department of Defense training and testing research and development community.

B. E. BACON
Captain, U.S. Navy
Commanding Officer

J. S. MCMICHAEL
Technical Director

SUMMARY

Problems

Many student assessment schemes currently used in Navy training are suspected of being insufficiently accurate or consistent. If true, this could result in either over-training, which increases costs needlessly, or undertraining, which culminates in unqualified graduates being sent to the fleets.

Objective

The specific objective of this research was to compare the reliability and validity of a computer-based and a paper-based procedure for assessing semantic knowledge.

Method

A Soviet threat-parameter database was compiled with the assistance of intelligence officers and instructors at VF-124, Naval Air Station (NAS) Miramar. This was structured as a semantic network in order to represent the associative knowledge inherent to it for the computer system. That is, objects and their corresponding properties, attributes, or characteristics were represented as node-link structures. The links between the nodes represent the associations or relationships among objects or among objects and their attributes.

A computer-based and paper-based test were designed and developed to assess this threat-parameter knowledge. Using a within-subjects experimental design, these tests were administered to 75 F-14 and E-2C crew members who volunteered to participate in this study. After subjects received one test, they were immediately given the other. It was assumed that a subject's state of threat-parameter knowledge was the same during the administration of both tests.

Reliabilities for both modes of testing were estimated by deriving internal consistency indices using an odd-even item split. These estimates were adjusted by employing the Spearman-Brown Prophecy Formula. Reliability estimates were calculated for test score, average degree of confidence, and average response latency for the computer-based test; reliability estimates were calculated for test score and average degree of confidence only for the paper-based test. None was computed for average response latency since this was not measured for the paper-based test. Equivalences between these two modes of assessment were estimated by Pearson product-moment correlations for total test score and average degree of confidence.

In order to derive discriminant validity estimates, research subjects were placed into groups according to three distinct grouping strategies: (a) above or below F-14 or E-2C mean flight hours, (b) F-14 radar intercept officers (RIOs) or pilots and E-2C naval flight officers (NFOs) or pilots, and (c) VF-124 students and instructors or members of other operational squadrons. Three stepwise multiple discriminant analyses, using Wilks' criterion for including and rejecting variables, and their associated statistics were computed to ascertain how well computer-based and paper-based

measures distinguished among the defined groups expected to differ in the extent of their knowledge of the threat-parameter database.

Results

This study established that (a) computer-based and paper-based measures, i.e., test score and average degree of confidence, are not significantly different in reliability or internal consistency; (b) for computer-based and paper-based measures, average degree of confidence has a higher reliability than average response latency which in turn has a higher reliability than the test score; (c) a few of the findings are ambivalent since some results suggest equivalence estimates for computer-based and paper-based measures, i.e., test score and average degree of confidence, are about the same, and another suggests these estimates are different; and (d) the discriminant validity of the computer-based measures was superior to paper-based measures.

Discussion and Conclusions

In this study, computer-based and paper-based testing were not significantly different in reliability with the former having more discriminant validity than the latter. These results suggest that computer-based assessment may have more utility for measuring semantic knowledge than paper-based measurement. This implies that the type of computerized testing used in this research may be better for estimating threat-parameter knowledge than traditional testing which has been primarily paper-based in nature.

The literature regarding computer-based assessment is contradictory and inconclusive: Many benefits may be obtained from computerized testing. Some of these may be related to attitudes and assumptions associated with the use of novel media or innovative technology per se. However, and just as readily, potential problems may result from the employment of computer-based measurement. Differences between this mode of assessment and traditional testing techniques may, or may not, impact upon the reliability and validity of measurement.

Recommendations

1. It is recommended that the computer-based test, FlashCards, be used to not only quiz but also train the threat-parameter database to F-14 and E-2C crew members. Currently, FlashCards and Jeopardy (the Computhreat system) are being used by VF-124 to augment the teaching and testing of threat parameters.

2. Other computer-based quizzes being developed at NPRDC should be used in different content areas to provide evidence about the generalizabiltiy of the reliability and validity findings established in this research.

CONTENTS

	Page
INTRODUCTION	1
Problems	1
Objective	1
METHOD	1
Subjects	1
Subject Matter	2
Computer-Based Assessment	2
Paper-Based Assessment	4
Procedure	4
RESULTS	5
Reliability and Equivalence Estimates	5
Discriminant Validity Estimates	6
Above or Below F-14 or E-2C Mean Flight Hours	6
F-14 PIOs or Pilots and E-2C NFOs or Pilots	7
VF-124 Students and Instructors or Members of Other Operational Squadrons	8
General Discriminant Validity	9
DISCUSSION AND CONCLUSIONS	9
RECOMMENDATIONS	12
REFERENCES	13
APPENDIX--TABLES OF RELIABILITY AND VALIDITY ESTIMATES	A-0
DISTRIBUTION LIST	

INTRODUCTION

Problems

Many student assessment schemes currently used in Navy training are suspected of being insufficiently accurate or consistent. If true, this could result in either over-training, which increases costs needlessly, or undertraining, which culminates in unqualified graduates being sent to the fleet commands. Many customary methods for measuring performance either on the job or in the classroom involve instruments which are primarily paper-based in nature (e.g., check lists, rating scales, critical incidences, and multiple-choice, completion, true-false, and matching formats). A number of deficiencies exist with these traditional testing techniques; e.g., (a) biased items are generated by different individuals, (b) item-writing procedures are usually obscure, (c) there is a lack of objective standards for producing tests, (d) item content is not typically sampled in a systematic manner, and (e) there is often a poor relationship between what is taught and test content.

What is required is a theoretically and empirically grounded technology of producing procedures for testing which will correct these faults. One promising approach employs computer technology. However, very few data are available regarding the psychometric properties of testing strategies using this technology. Data are needed concerning the accuracy, consistency, sensitivity, and fidelity of these computer-based assessment schemes compared to more traditional testing techniques.

Objective

The specific objective of this research was to compare the reliability and validity of a computer-based and a paper-based procedure for assessing semantic knowledge.

METHOD

Subjects

The subjects were 75 F-14 pilots, radar intercept officers (RIOs), and students as well as E-2C pilots and naval flight officers (NFOs) from operational squadrons at Naval Air Station (NAS) Miramar who had volunteered to participate in this research. The primary test-bed has been the Fleet Replacement Squadron, VF-124, NAS Miramar. The main reason this squadron exists is to train pilots and RIOs for the F-14 fighter. One of the major missions of the F-14 is to protect carrier-based naval task forces against antiship, missile-launching, threat bombers. This part of the F-14's mission is referred to as Maritime Air Superiority (MAS), which is taught in the Advanced Fighter Air Superiority (ADFAS) curriculum in the squadron. It is during ADFAS that the students must learn a threat-parameter database so that they can properly employ the F-14 against hostile platforms. E-2C pilots, NFOs, and students receive similar instruction. The tests currently administered to these officers are primarily paper-based in nature and normally formatted as multiple choice and

completion items.

Subject Matter

A classified database was developed consisting of five categories of facts about front-line Soviet platforms: weapons systems, radar and ECM systems, surface and subsurface platforms, airborne platforms, and counterjamming procedures. It was used to train and test F-14 pilots, RIOs, and students concerning important threat parameters associated with Russian platforms: e.g., aircraft range and speed, payload of antiship missiles, typical launch altitude; missile range, flight profile, velocity, and warheads; other weapon, radar, electronic countermeasure (ECM)/ electronic counter-countermeasure (ECCM) systems; and surveillance capabilities.

The database was compiled with the assistance of the intelligence officers and the ADFAS instructors of VF-124. It was structured as a semantic network (Barr & Feigenbaum, 1981; Johnson-Laird, 1983) in order to represent the associative knowledge inherent to it for the computer system. That is, objects and their corresponding properties, attributes, or characteristics were represented as node-link structures. The links between those nodes represent the associations or relationships among objects or among objects and their attributes. For example, the object "aircraft type" and the attribute "ECM suite" can be linked so that the system can represent a particular aircraft type that has a certain ECM suite. By defining initially all objects and attributes in the database, a hierarchy or tree structure can be specified for all objects, attributes, and their relationships. A typical database can contain representations of several thousands of such associations. The database can also include synonyms and quantifiers. The former allows an object to be specified or referred to in several ways; the latter allows the number of certain attributes to be associated with a particular object.

Computer-Based Assessment

Once a database was structured as a semantic network, it became possible for independent software modules to interact with, operate upon, or manipulate the database. For example, interpretative programs could make inferences about the subject database, or they could ask questions about the database since its intrinsic structure was represented. This latter capability was capitalized upon in this research.

A computer-based game was adopted and adapted to quiz students and instructors in VF-124 as well as crew members of other operational squadrons that belong to the wing at NAS Miramar about the threat-parameter database. This computer-based quiz, or test, is totally independent of the database and will run on any database structured as a semantic network. It will randomly select objects from the database, and generate questions about them and their attributes. Unlike some computer-based tests, alternative forms did not have to be specifically programmed as such.

With the database represented as a semantic network, it was feasible to employ one of the games or quizzes that was programmed as a component of the Computer-Based Tactical Memorization Training System developed by the Navy Personnel Research and Development Center (NPRDC) under the work unit entilted: Computer-

Based Techniques for Training Tactical Knowledge, RF63-522-801-013.03.02. To reiterate, the games are autonomous entities which can operate on any database that can be structured as a semantic network. These games can quiz students by randomly choosing characteristics or objects from the database, and generating questions about threat platforms and their salient attributes.

One of the computer-based games that was chosen from this prior NPRDC development for conducting this research is called FlashCards. It was substantially improved to yield: more experimental control, measures of response latencies and degrees of confidence in responses, and better record keeping for assessing student performance, facilitating the computation of statistical analyses, and presenting feedback to the instructors and students. These programming enhancements were documented by Liggett and Federico (1986). The computer-based system containing FlashCards and another game, Jeopardy together with the threat-parameter database for the F-14 and E-2C communities is referred to as Computhreat.

FlashCards is analogous to using real flash cards. That is, a question is presented to individual students who are expected to answer it. Questions can have multiple answers as in "What Soviet bombers carry the XYZ-123 missile?" After individual students are presented with the question, they are allowed as many tries as they would like to answer. If the students cannot answer the question, they can continue with the game. At this point, they are presented with the correct answer or answers. At any point in the answering process, they can continue to the next question. For each answer, the students must key in a response which reflects their degree of confidence in their answer. Also, for each answer, the student's response latency is recorded and displayed.

FlashCards will quiz the students on all top-level, or general, categories of the semantic network that it is using as the database. After the game, students are given feedback as to their overall performance. FlashCards keeps records of a student's: latency, confidence, overall score, number answered correctly, number answered incorrectly, and number not answered. Records are kept across all items for each student.

A question cycle begins with an individual student being prompted with a question and the number of correct answers required to fully answer that question. Also visible is an empty Correct Answers Menu which is a box structure that will hold all the correct answers. An answer will be placed there when an individual answers a question correctly, or gives up in which case the program divulges the correct answer(s). The testee is notified that a clock has started, and is then required to type in an answer. After typing <return> at the end of the answer, the individual is given response time in seconds, and presented with a scale ranging from zero to one-hundred percent in ten point intervals to be used to indicate the percentage of confidence or the degree of sureness the testee has in the answer(s). The student is then required to type in a single digit corresponding to the selected confidence level. After the confidence value is entered, the testee is notified if the answer was correct or incorrect. If correct, the answer is put into the Correct Answers Menu and the number of answers left to be entered is decremented. If that number is zero, the question terminates and program control is passed to the next question. If the answer is incorrect, the individual is merely prompted again to enter an answer. If the testee does not know all the correct

answers, A <return> may be entered to put all the remaining correct answers in the Correct Answers Menu.

The score for each question was computed as the number of correct answers entered divided by the total number of answers entered. A <return> was not counted as an answer. For the purposes of this research, a complete FlashCards test consisted of 25 domain-referenced items or questions. These were considered as two groups of 12 odd and even items each, dropping the last question, for computing split-half reliability estimates. The average score for odd (even) items was calculated as the total score of odd (even) items divided by the number of odd (even) questions attempted. The total computer-based test score was calculated as the average of the odd and even halves.

The software for the complete gaming system is currently on eight floppy disks. The game itself is run with only two dual-density disks on a Terak microcomputer employing two drives. It is implemented on the UCSD P-system and written in UCSD-Pascal. The disk placed in the bottom drive holds the actual game code; the disk placed in the top drive contains the independent semantic network database. As soon as the system is booted, control is passed to the game. Consequently, naive users need not deal with the nuances of the UCSD P-system. Knowledge-performance data for the FlashCards game are saved for individual players on the disk in the lower drive. There are six other disks that contain files necessary for modification of the gaming system and/or data collection. These disks contain the text of the games, the semantic network database, the statistical programs, and all necessary P-system files.

Paper-Based Assessment

Two alternative forms of a paper-based test were designed and developed to assess knowledge of the same threat-parameter database mentioned above, and to mimick as much as possible the format used by FlashCards. Both of these consisted of 25 completion or fill-in-the-blank domain-referenced items. As with the computer-based test, more than one answer may be required per item or question. Beneath each question was a confidence scale which resembled the one used in FlashCards where the testees were required to indicate the level of confidence in their response(s). Scoring items for this paper-based test was similar to scoring the computer-based test: For each question, the number of correct answers given was divided by the total number of answers completed for that question. Also, scoring odd (even) halves of the test for computing internal consistency was similar to that for FlashCards. The score for the total paper-based test was calculated like the total score for the computer-based test.

Procedure

Subjects acquired threat-parameter knowledge using dual media: (1) a traditional text organized according to the database's major topics, and (2) the Computhreat computer-based system. Mode of assessment, computer-based or paper-based, was manipulated as a within-subjects variable. Subjects were administered the computer-based and paper-based tests in counterbalanced order. The two forms of the paper-based tests were alternated in their administration to subjects, i.e., the first subject

received Form A, the second subject received Form B, the third subject received Form A, etc. After subjects received one test, they were immediately administered the other. It was assumed that a subject's state of threat-parameter knowledge was the same during the administration of both tests. Subjects took approximately 10-15 minutes to complete the paper-based test, and 20-25 minutes to complete the computer-based test. The longer time to complete the latter test was largely attributed to lack of typing or keyboard proficiency on the part of some of the subjects.

Reliabilities for both modes of testing were estimated by deriving internal consistency indices using an odd-even item split. These reliability estimates were adjusted by employing the Spearman-Brown Prophecy Formula (Thorndike, 1982). Reliability estimates were calculated only for test score, average degree of confidence, and average response latency for the computer-based test; reliability estimates were calculated for test score and average degree of confidence for the paper-based test. None was computed for average response latency since this was not measured for the paper-based test. Equivalences between the two modes of assessment were estimated by Pearson product-moment correlations for total test score and average degree of confidence. These correlations were considered indices of the extent to which the two types of testing were measuring the same semantic knowledge and amount of assurance in answers.

In order to derive discriminant validity estimates, research subjects were placed into groups according to three distinct grouping strategies: (a) above or below F-14 or E-2C mean flight hours, (b) F-14 RIOs or pilots and E-2C NFOs or pilots, and (c) VF-124 students and instructors or members of other operational squadrons. Three stepwise multiple discriminant analyses, using Wilks' criterion for including and rejecting variables, and their associated statistics were computed to ascertain how well computer-based and paper-based measures distinguished among the defined groups expected to differ in the extent of their knowledge of the threat-parameter database. It was thought that mean flight hours reflect operational experience. Those individuals with more operational experience were expected to perform better on tests of threat-parameter knowledge than those with less experience. It was thought that F-14 crew members would have knowledge superior to E-2C crew members regarding threat parameters because of the difference in their operational missions and training emphasis. Lastly, it was expected that students would do better on tests of threat-parameter knowledge because their exposure to this subject matter was more recent to that of instructors and members of other operational crews who probably had not reviewed this material for sometime.

RESULTS

Reliability and Equivalence Estimates

Tables of reliability and validity estimates are presented in the appendix. Split-half reliability and equivalence estimates of computer-based and paper-based measures from the pooled within-groups correlation matrices for the different groupings are tabulated in Table 1. It can be seen that the adjusted reliability estimates of the computer-

based and paper-based measures are from moderate to high for the different groupings ranging from: (a) .73 to .97 for F-14 RIO and pilot and E-2C NFO and pilot, (b) .74 to .97 for above and below mean flight hours, and (c) .53 to .95 for student, instructor, and other. None of the differences in corresponding reliabilities for computer-based and paper-based measures, i.e., test score and average degree of confidence, were found to be statistically significant ($p > .01$) using a test described by Edwards (1964). This suggested that the computer-based and paper-based measures were not significantly different in reliability or internally consistency.

Considering the computer-based measures for all groupings, it was ascertained that the reliability estimate for average degree of confidence was significantly ($p < .01$) higher than the reliability estimates for average response latency and test score. Also, the reliability estimate for response latency was significantly higher than the one computed for test score. Focusing on the paper-based measures for all groupings, it was found that the reliability estimate for average degree of confidence was significantly ($p < .01$) higher than the reliability estimate for test score. These results implied that these measures can be ranked in order of their internal consistencies from highest to lowest as follows: average degree of confidence, average response latency, and test score.

Equivalence estimates for the different groupings reported in the same order as above for test score and average degree of confidence measures, respectively, were .76 and .82, .76 and .82, and .50 and .76. These suggested that the computer-based and paper-based measures had anywhere from 25% to 67% variance in common implying that these different modes of assessment were somewhat or partially equivalent. Equivalence is somewhat limited by the low reliability obtained for the computer-based measure of test score for the grouping: students, instructors, or others. For the F-14/E-2C and mean flight hours groupings, the equivalences for test score and average degree of confidence measures were not significantly ($p > .01$) different. However, for the student/instructor grouping, the equivalences of these measures were found to be significantly ($p < .01$) different. These results are ambiguous in that some of them suggest that the equivalence estimates for test score and average degree of confidence measures are about the same; while, the other suggests that these estimates are different.

Discriminant Validity Estimates

Above or Below F-14 or E-2C Mean Flight Hours

The discriminant analysis computed to determine how well computer-based and paper-based measures differentiated groups defined by above or below F-14 or E-2C mean flight hours yielded one significant discriminant function. According to the multiple discriminant analysis model (Cooley & Lohnes, 1962; Tatsuoka, 1971; Van de Geer, 1971), the maximum number of derived discriminant functions is either one less than the number of groups or equal to the number of discriminating variables, whichever is smaller. Since there were four groups to be discriminated, this analysis yielded three discriminant functions, but only one of them was significant. Consequently, solely this significant discriminant function and its associated statistics are presented.

The statistics associated with the significant function, standardized discriminant-function coefficients, pooled within-groups correlations between the function and computer-based and paper-based measures, and group centroids for above or below F-14 or E-2C mean flight hours are presented in Table 2. It can be seen that the single significant discriminant function accounted for approximately 82% of the variance among the four groups. The discriminant-function coefficients which consider the interactions among the multivariate measures revealed the relative contribution or comparative importance of these variables in defining this derived dimension to be the paper-based test total score (PTS), the computer-based test total score (CTS), and the computer-based test total average degree of confidence (CTC), respectively. The computer-based test total average latency (CTL) and the paper-based test total average degree of confidence (PTC) were considered unimportant in specifying this discriminant function since the absolute value of their coefficients were each below .4. The within-groups correlations which are computed for each individual measure partialling out the interactive effects of all the other variables indicated that the major contributors to the significant discriminant function were CTC, CTS, and CTL, respectively, all computer-based measures. The group centroids showed how the performance of the F-14 crew members clustered together along one end of the derived dimension; while, the performance of the E-2C crew members clustered together along the other end of the continuum. The means and standard deviations for groups above or below F-14 or E-2C mean flight hours, univariate F-ratios, and levels of significance for computer-based and paper-based measures are tabulated in Table 3. Considering the measures as univariate variables, i.e., independent of their multivariate relationships with one another, these statistics revealed that the three computer-based measures CTC, CTS, and CTL, respectively, significantly differentiated the four groups, not the paper-based measures, PTS and PTC. Applying Duncan's multiple range test (Kirk, 1968) on the group means for the important individual measures indicated that F-14 crews significantly ($p < .05$) out performed E-2C crews on CTS, CTC, and CTL. The multivariate and subsequent univariate results established the discriminant validity of computer-based measures to be superior to that of paper-based measures for the grouping strategy: above or below F-14 or E-2C flight hours.

F-14 RIOs or Pilots and E-2C NFOs or Pilots

The statistics associated with the significant function, standardized discriminant function coefficients, pooled within-groups correlations between the function and computer-based and paper-based measures, and group centroids for F-14 RIOs or pilots and E-2C NFOs or pilots are presented in Table 4. A single significant discriminant function accounted for approximately 82% of the variance among the four groups. The discriminant-function coefficients revealed the relative contribution of the multivariate measures in defining this derived dimension to be PTS, CTS, CTL, and PTC, respectively. CTC was considered unimportant in specifying this discriminant function since the absolute value of its coefficient was below .4. The within-groups correlations for the measures indicated that the major contributors to the significant discriminant function were CTC, CTS, CTL, and PTC, respectively. Seventy-five percentage of these were computer-based measures. The group centroids showed how the performance of the F-14 crew members clustered together along one end of the derived dimension; while, the performance of the E-2C crew members was spread out along the other end

of the continuum. The means and standard deviations for groups of F-14 RIOs or pilots and E-2C NFOs or pilots, univariate F-ratios, and levels of significance for computer-based and paper-based measures are tabulated in Table 5. Considering the measures as univariate variables, these statistics revealed that the three computer-based measures CTL, CTS, CTC, and one paper-based measure, PTC, respectively, significantly differentiated the four groups. Applying Duncan's multiple range test on the group means for these individual measures indicated that (a) F-14 crews significantly ($p < .05$) out performed E-2C crews on CTS and CTC; and (b) F-14 crew members and E-2C NFOs significantly out performed E-2C pilots on CTL and PTC measures. The multivariate and univariate results established the discriminant validity of the computer-based measures to be greater than the paper-based measures for the grouping strategy: F-14 RIOs or pilots and E-2C NFOs or pilots.

VF-124 Students and Instructors or Members of Other Operational Squadrons

The statistics associated with the significant function, standardized discriminant-function coefficients, pooled within-groups correlations between the function and computer-based and paper-based measures, and group centroids for VF-124 students and instructors or members of other operational squadrons are presented in Table 6. A single significant discriminant function accounted for approximately 98% of the variance among the three groups. The discriminant-function coefficients revealed the relative contribution of the multivariate measures in defining this derived dimension to be CTS and CTC, respectively. The within-groups correlations for the measures indicated that the major contributors to the significant discriminant function were CTS, CTC, PTS, and PTC, respectively. Half of these were computer-based measures, and half were paper-based measures. The group centroids showed how the performances of the students, instructors, and others were spread out along the entire dimension. The means and standard deviations for groups of VF-124 students and instructors or members of other operational squadrons, univariate F-ratios, and levels of significance for computer-based and paper-based measures are tabulated in Table 7. Considering the measures as univariate variables, these statistics revealed that all three computer-based measures CTS, CTC, CTL, and the two paper-based measures, PTS and PTC, respectively, significantly differentiated the three groups. Applying Duncan's multiple range test on the group means for these individual measures indicated that (a) students significantly ($p < .05$) out performed instructors who in turn did better than members of other operational squadrons on CTS; (b) students and instructors did equally well but significantly out performed members of other operational squadrons on CTC, CTL, and PTC; and (c) students did significantly better than instructors and others who performed equally well on PTS. The multivariate and univariate results established the discriminant validity of the computer-based measures to be higher than paper-based measures for the grouping strategy: VF-124 students and instructors or members of other operational squadrons.

General Discriminant Validity

Distinguishing among the groups formed by the three grouping strategies suggested that, generally, the discriminant validity of the computer-based measures was superior to that of the paper-based measures.

Discussion and Conclusions

This study established that (a) computer-based and paper-based measures, i.e., test score and average degree of confidence, are not significantly different in reliability or internal consistency; (b) for computer-based and paper-based measures, average degree of confidence has a higher reliability than average response latency which in turn has a higher reliability than the test score; (c) a few of the findings are ambivalent since some results suggest equivalence estimates for computer-based and paper-based measures, i.e., test score and average degree of confidence, are about the same, and another suggests these estimates are different; and (d) the discriminant validity of the computer-based measures was superior to paper-based measures. The results of this research supported the findings of some studies, but not others. The reported literature on this subject is contradictory and inconclusive.

The consequences of computer-based assessment on examinees' performance are not obvious. The few studies that have been conducted on this topic have produced mixed results. Investigations of computer-based administration of personality items have yielded reliability and validity indices comparable to typical paper-based administration (Katz & Dalby, 1981; Lushene, O'Neil, & Dunn, 1974). No significant differences were found in the scores of measures of anxiety, depression, and psychological reactance due to computer-based and paper-based administration (Lukin, Dowd, Plake, & Kraft, 1985). Studies of cognitive tests have provided inconsistent findings with some (Rock & Nolen, 1982; Hitti, Riffer, & Stuckless, 1971) demonstrating that the computerized version is a viable alternative to the paper-based version. Other research (Hansen & O'Neil, 1970; Hedl, O'Neil, & Hansen, 1973; Johnson & White, 1980; Johnson & Johnson, 1981), though, indicated that interacting with a computer-based system to take an intelligence test could elicit a considerable amount of anxiety which could affect performance.

Some studies (Serwer & Stolurow, 1970; Johnson & Mihal, 1973) demonstrated that testees do better on verbal items given by computer than paper-based; however, just the opposite was found by other studies (Johnson & Mihal, 1973; Wildgrube, 1982). One investigation (Sachar & Fletcher, 1978) yielded no significant differences resulting from computer-based and paper-based modes of administration on verbal items. Two studies (English, Reckase, & Patience, 1977; Hoffman & Lundberg, 1976) demonstrated that these two testing modes did not affect performance on memory retrieval items. Sometimes (Johnson & Mihal, 1973) testees performed better on quantitative tests when computer given; sometimes (Lee, Moreno, & Sympson, 1984) they performed worse; and other times (Wildgrube, 1982) it may make no difference. Other studies have supported the equivalence of computer-based and paper-and-paper administration (Elwood & Griffin, 1972; Hedl, O'Neil, & Hansen, 1973; Kantor, 1988; Lukin, Dowd, Plake, & Kraft, 1985). Some researchers (Evan & Miller, 1969; Koson, Kitchen, Kochen, & Stodolosky, 1970; Lucas, Mullin, Luna, & McInroy, 1977; Lukin,

Dowd, Plake, & Kraft, 1985; Skinner & Allen, 1983) have reported comparable or superior psychometric capabilities of computer-based assessment relative to paper-based assessment in clinical settings.

Regarding computerized adaptive testing (CAT), some empirical comparisons (McBride, 1980; Sympson, Weiss, & Ree, 1982) yielded essentially no change in validity due mode of administration. However, test item difficulty may not be indifferent to manner of presentation for CAT (Green, Bock, Humphreys, Linn, & Reckase, 1984). When going from paper-based to computer-based administration, this mode effect is thought to have three aspects: (a) an overall mean shift where all items may be easier or harder, (b) an item mode interaction where a few items may be altered and others not, and (c) the nature of the task itself may be changed by computer administration. A computer simulation study (Divgi, 1988) demonstrated that a CAT version of the Armed Services Vocational Aptitude Battery had higher reliability than a paper-based version for these subtests: General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, and Mathematics Knowledge. These inconsistent results of mode, manner, or medium of testing may be due to differences in methodology, test content, population tested, or the design of the study (Lee, Moreno & Sympson, 1984).

With computer costs coming down and peoples' knowledge of these systems going up, it becomes more likely economically and technologically that many benefits can be gained from their use. Some indirect advantages of computer-based assessment are increased test security, less ambiguity about students' responses, minimal or no paperwork, immediate scoring, and automatic records keeping for item analysis (Green, 1983a, 1983b). Some of the strongest support for computer-based assessment is based upon the awareness of faster and more economical measurement (Elwood & Griffin, 1972; Johnson & White, 1980; Space, 1981). Cory (1977) reported some advantages of computerized over paper-based testing for predicting on job performance.

Ward (1984) stated that computers can be employed to augment what is possible with paper-based measurement, e.g, to obtain more precise information regarding a student than is likely with more customary measurement methods, and to assess additional aspects of performance. He discussed potential benefits that may be derived from employing computer-based systems to administer traditional tests. Some of these are as follows: (a) individualizing assessment, (b) increasing the flexibility and efficiency for managing test information, (c) enhancing the economic value and manipulation of measurement databases, and (d) improving diagnostic testing. Millman (1984) claimed to agree with Ward, especially regarding the ideas that computer-based measurement encourages: individualizing assessment, designing software within the context of cognitive science, and limiting computer-based assessment is not hardware inadequacy but incomplete comprehension of the processes intrinsic to testing and knowing per se (Federico, 1980).

Sampson (1983) discussed some of the potential problems associated with computer-based assessment: (a) not taking into account human factors principles to design the human-computer interface, (b) individuals becoming so anxious when interacting with a computer for assessment that the measurement obtained may be questionable, (c) possibility of unauthorized access and invasion of privacy, (d) inaccurate test interpretations by users of the system culminate in erroneously drawn

conclusions, (e) differences in modes of administration making paper-based norms inappropriate for computer-based assessment, (f) lack of reporting reliability and validity data for computerized tests, and (g) resistance toward using new computer-based systems for performance assessment. A potential limitation of computer-based assessment is depersonalization and decreased opportunity for observation. This is especially true in clinical environments (Space, 1981). Most computer-based tests do not allow individuals to omit or skip items, or to alter earlier responses. This procedure could change the test-taking strategy of some examinees. To permit it, however, would probably create confusion and hesitation during the process of retracing through items as the testee uses clues from some to minimize the degree of difficulty of others (Green, Bock, Humphreys, Linn, & Reckase, 1984).

Hofer and Green (1985) were concerned that computer-based assessment would introduce irrelevant or extraneous factors that would likely degrade test performance. These computer-correlated factors may alter the nature of the task to such a degree, it would be difficult for a computer-based test and its paper-based counterpart to measure the same construct or content. This could impact upon reliability, validity, normative data, as well as other assessment attributes. They listed several factors which might contribute to different performances on these distinct kinds of testing: (a) state anxiety instigated when confronted by computer-based testing, (b) lack of computer familiarity on the part of the testee, and (c) changes in response format required by the two modes of assessment. These different dimensions could result in tests that are nonequivalent; however, in this reported research, these diverse factors had no apparent impact.

A number of known differences between computer-based and paper-based assessment which may affect equivalence and validity are as follows: No passive omitting of items is usually permitted on computer-based tests. An individual must respond unlike most paper-based tests. Computerized tests typically do not permit backtracking. The testee cannot easily review items, alter responses, or delay attempting to answer questions. The capacity of the computer screen can have an impact on what usually are long test items, e.g., paragraph comprehension. These may be shortened to accommodate the computer display, thus partially changing the nature of the task. The quality of computer graphics may affect the comprehension and degree of difficulty of the item. Pressing a key or using a mouse is probably easier than marking an answer sheet. This may impact upon the validity of speeded tests. Since the computer typically displays items individually, traditional time limits are no longer necessary. The multidimensionality of achievement tests has implications for scoring CATs (Green, 1986).

Some of the comments made by Colvin and Clark (1984) concerning instructional media can easily be extrapolated to assessment media. (Training and testing are inextricably intertwined; it is difficult to do one well without the other.) This is especially appropriate regarding some of the attitudes and assumptions permeating the employment of, and enthusiasm for, media: (a) confronted with new media, computer-based or otherwise, students will not only work harder, but also enjoy their training and testing more; (b) matching training and testing content to mode of presentation is important, even though not all that prescriptive or empirically well established; (c) the application of computer-based systems permits self-instruction and self-assessment with their concomitant flexibility in scheduling and pacing training and testing; (d) monetary and

human resources can be invested in designing and developing computer-based media for instruction and assessment that can be used repeatedly and amortized over a longer time, rather than in labor intensive classroom-based training and testing; and (e) the stability and consistency of instruction and assessment can be improved by media, computer-based or not, for distribution at different times and locations however remote.

Evaluating or comparing different media for instruction and assessment, one must be aware that the newer medium may simply be perceived as being more novel, interesting, engaging, and challenging by the students. This novelty effect seems to disappear as rapidly as it appears. However; in research studies conducted over a relatively short time span, e.g., a few days or months at the most, this effect may still be lingering and affecting the evaluation by enhancing the impact of the more novel medium (Colvin & Clark, 1984). When matching media to distinct subject matters, course contents, or core concepts, some research evidence (Jamison, Suppes, & Welles, 1974) indicates that, other than in obvious cases, just about any medium will be effective for different content.

As is evident, the literature regarding computer-based assessment is contradictory and inconclusive: Many benefits may be obtained from computerized testing. Some of these may be related to attitudes and assumptions associated with the use of novel media or innovative technology per se. However, and just as readily, potential problems may result from the employment of computer-based measurement. Differences between this mode of assessment and traditional testing techniques may, or may not, impact upon the reliability and validity of measurement.

In this study, it was found that computer-based and paper-based testing were not significantly different in reliability with the former having more discriminant validity than the latter. These results suggest that computer-based assessment may have more utility for measuring semantic knowledge than paper-based measurement. This implies that the type of computerized testing used in this research may be better for estimating threat-parameter knowledge than traditional testing which has been primarily paper-based in nature.

A salient question that needs to be addressed is how to combine effectively and efficiently computer and cognitive science, artificial intelligence (AI), current psychometric theory, and diagnostic testing. AI techniques can be developed to diagnose specific error-response patterns or bugs to advance measurement methodology (Brown & Burton, 1978; Kieras, 1987; McArthur & Choppin, 1984).

Recommendations

1. It is recommended that the computer-based test, FlashCards, be used to not only quiz but also train the threat-parameter database to F-14 and E-2C crew members. Currently, FlashCards and Jeopardy (the Computhreat system) are being used by VF-124 to augment the teaching and testing of threat parameters.

2. Other computer-based quizzes being developed at NPRDC should be used in different content areas to provide evidence on the generalizabiltiy of the reliability and validity findings established in this research.

References

Barr, A., & Feigenbaum, E. F. (Eds.). (1981). *The handbook of artificial intelligence, Volume 1.* Stanford CA: HeurisTech.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in mathematical skills. *Cognitive Science, 2,* 155-192.

Colvin, C., & Clark, R. E. (1984). Instructional media vs. instructional methods. *Performance and Instruction Journal,* July, 1-3.

Cooley, W. W., & Lohnes, P. R. (1962). *Multivariate procedures for the behavioral sciences.* New York: John Wiley & Sons.

Cory, C. H. (1977). Relative utility of computerized versus paper-and-pencil tests for predicting job performance. *Applied Psychological Measurement, 1,* 551-564.

Divgi, D. R. (1988, October). *Two consequences of improving a test battery* (CRM 88-171). Alexandria VA: Center for Naval Analyses.

Edwards, A. L. (1964). *Experimental design in psychological research.* New York: Holt, Rinehart, and Winston.

Elwood, D. L., & Griffin, R. H. (1972). Individual intelligence testing without the examiner: Reliability of an automated method. *Journal of Consulting and Clinical Psychology, 38,* 9-14.

English, R. A., Reckase, M. D., & Patience, W. M. (1977). Applications of tailored testing to achievement measurement. *Behavior Research Methods & Instrumentation, 9,* 158-161.

Evan, W. M., & Miller, J. R. (1969). Differential effects of response bias of computer versus conventional administration of a social science questionnaire. *Behavioral Science, 14,* 216-227.

Federico, P-A. (1980). Adaptive instruction: Trends and issues. In R. E. Snow, P-A. Federico, & W. E. Montague (Eds.), *Aptitude, learning, and instruction, Volume 1: Cognitive process analyses of aptitude.* Hillsdale NJ: Erlbaum.

Green, B. F. (1983a). Adaptive testing by computer. *Measurement, Technology, and Individuality in Education. 17,* 5-12.

Green, B. F. (1983b). The promise of tailored tests. In H. Wainer & S. Messick (Eds.) *Principles of modern psychological measurement: A festschrift in honor of Frederic Lord.* Hillsdale NJ: Erlbaum.

Green, B. F. (1986). *Construct validity of computer-based tests.* Paper presented at the test validity conference educational testing service, Princeton, N. J.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347-360.

Hansen, D. H., & O'Neil, H. F. (1970). Empirical investigations versus anecdotal observations concerning anxiety and computer-assisted instruction. *Journal of School Psychology, 8,* 315-316.

Hedl, J. J., O'Neil, H. F., & Hansen, D. H. (1973). Affective reactions toward computer-based intelligence testing. *Journal of Consulting and Clinical Psychology, 40,* 217-222.

Hitti, F. J., Riffer, R. L., & Stuckless, E. R. (1971, July). *Computer-managed testing: A feasibility study with deaf students.* National Technical Institute for the Deaf.

Hofer, P. J., & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology, 53,* 826-838.

Hoffman, K. I., & Lundberg, G. D. (1976). A comparison of computer-monitored group tests with paper-and-pencil tests. *Educational and Psychological Measurement, 36,* 791-809.

Jamison, D., Suppes, P., & Welles, S. (1974). The effectiveness of alternative media: A survey. *Annual Review of Educational Research, 44,* 1-68.

Johnson, J. H., & Johnson, K. N. (1981). Psychological considerations related to the development of computerized testing stations. *Behavior Research Methods & Instrumentation, 13,* 421-424.

Johnson, D. F., & Mihal, W. L. (1973). Performance of black and whites in computerized versus manual testing environments. *American Psychologist, 28,* 694-699.

Johnson, D. F., & White, C. B. (1980). Effects of training on computerized test performance in the elderly. *Journal of Applied Psychology, 65,* 357-358.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge MA: Harvard University Press.

Kantor, J. (1988). *The effects of anonymity, item sensitivity, trust, and method of administration on response bias on the job description index.* Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Katz, L., & Dalby, J. T. (1981). Computer-assisted and traditional psychological assessment of elementary-school-age children. *Contemporary Educational Psychology, 6,* 314-322.

Kieras, D. E. (1987). *The role of cognitive simulation models in the development of advanced training and testing systems* (TR-87/ONR-23). Ann Arbor: University of Michigan.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences.* Belmont CA: Brooks/Cole.

Koson, D., Kitchen, C., Kochen, M., & Stodolosky, D. (1970). Psychological testing by computer: Effect on response bias. *Educational and Psychological Measurement, 30,* 808-810.

Lee, J. A., Moreno, K. E., & Sympson, J. B. (1984, April). *The effects of mode of test administration on test performance.* Paper presented at the annual meeting of the Eastern Psychological Association, Baltimore.

Liggett, N. L., & Federico, P-A. (1986). *Computer-based system for assessing semantic knowledge: Enhancements* (NPRDC TN 87-4). San Diego: Navy Personnel Research and Development Center.

Lucas, R. W., Mullin, P. J., Luna, C. D., & McInroy, D. C. (1977). Psychiatrists and a computer as interrogators of patients with alcohol related illnesses: A comparison. *British Journal of Psychiatry, 131,* 160-167.

Lukin, M. E., Dowd, E. T., Plake, B. S., & Kraft, R. G. (1985). Comparing computerized versus traditional psychological assessment. *Computers in Human Behavior, 1,* 49-58.

Lushene, R. E., O'Neil, H. F., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment, 34,* 353-361.

McArthur, D. L., & Choppin, B. H. (1984). Computerized diagnostic testing. *Journal*

of Educational Measurement, 21, 391-397.

McBride, J. R. (1980). Adaptive verbal ability testing in a military setting. In D. J. Weiss (Ed.), *Proceedings of the 1979 computerized adaptive testing conference.* Minneapolis: University of Minnesota, Department of Psychology.

Millman, J. (1984). Using microcomputers to administer tests: An alternate point of view. *Educational Measurement: Issues and Practices,* Summer, 20-21.

Rock, D. L., & Nolen, P. A. (1982). Comparison of the standard and computerized versions of the raven coloured progressive matrices test. *Perceptual and Motor Skills, 54,* 40-42.

Sachar, J. D., & Fletcher, J. D. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology.

Sampson, J. R. (1983). Computer-assisted testing and assessment: Current status and implications for the future. *Measurement and Evaluation in Guidance, 15,* 293-299.

Serwer, B. L., & Stolurow, L. M. (1970). Computer-assisted learning in language arts. *Elementary English, 47,* 641-650.

Skinner, H. A., & Allen, B. A. (1983). Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. *Journal of Consulting and Clinical Psychology, 51,* 267-275.

Space, L. G. (1981). The computer as psychometrician. *Behavior Research Methods & Instrumentation, 13,* 595-606.

Sympson, J. B., Weiss, D. J., & Ree, M. (1982). *Predictive validity of conventional and adaptive tests in an air force training environment* (AFHRL-TR-81-40). Brooks AFB: Air Force Human Resources Laboratory.

Tatsuoka, M. M. (1971). *Multivariate analysis.* New York: John Wiley & Sons.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

Van de Geer, J. P. (1971). *Introduction to multivariate analysis for the social sciences.* San Francisco: W. H. Freeman.

Ward, W. C. (1984). Using microcomputers to administer tests. *Educational Measurement: Issues and Practices,* Summer, 16-20.

Wildgrube, W. (1982, July). *Computerized testing in the german federal armed forces--empirical approaches.* Paper presented at the 1982 Computerized Adaptive Testing Conference, Spring Hill MN.

APPENDIX

TABLES OF RELIABILITY AND VALIDITY ESTIMATES

Page

1. Split-Half Reliability and Equivalence Estimates of Computer-Based and Paper-and-Pencil Measures from Pooled Within-Groups Correlation Matrices for Different Groupings A-1

2. Statistics Associated with Significant Discriminant Function, Standardized Discriminant-Function Coefficients, Pooled Within-Groups Correlations Between the Discriminant Function and Computer-Based and Paper-and-Pencil Measures, and Group Centroids for Above or Below F-14 or E-2C Mean Flight Hours A-2

3. Means and Standard Deviations for Groups Above or Below F-14 or E-2C Mean Flight Hours, Univariate F-Ratios, and Levels of Significance for Computer-Based and Paper-and-Pencil Measures A-3

4. Statistics Associated with Significant Discriminant Function, Standardized Discriminant-Function Coefficients, Pooled Within-Groups Correlations Between the Discriminant Function and Computer-Based and Paper-and-Pencil Measures, and Group Centroids for F-14 RIOs or Pilots and E-2C NFOs or Pilots A-4

5. Means and Standard Deviations for Groups of F-14 RIOs or Pilots and E-2C NFOs or Pilots, Univariate F-Ratios, and Levels of Significance for Computer-Based and Paper-and-Pencil Measures A-5

6. Statistics Associated with Significant Discriminant Function, Standardized Discriminant-Function Coefficients, Pooled Within-Groups Correlations Between the Discriminant Function and Computer-Based and Paper-and-Pencil Measures, and Group Centroids for VF-124 Students and Instructors or Members of Other Operational Squadrons A-6

7. Means and Standard Deviations for Groups of VF-124 Students and Instructors or Members of Other Operational Squadrons, Univariate F-Ratios, and Levels of Significance for Computer-Based and Paper-and-Pencil Measures A-7

Table 1

Split-Half Reliability and Equivalence Estimates of Computer-Based and Paper-and-Pencil Measures from Pooled Within-Groups Correlation Matrices for Different Groupings

Grouping	Above or Below Mean Flight Hours		
	Reliability		Equiva-
Measure	Computer-Based	Paper-and-Pencil	lence
Score	.74	.76	.76
Confidence	.96	.97	.82
Latency	.88	-	-
Grouping	F-14 RIOs/Pilots, E2-C NFOs/Pilots		
	Reliability		Equiva-
Measure	Computer-Based	Paper-and-Pencil	lence
Score	.73	.77	.76
Confidence	.95	.97	.82
Latency	.86	-	-
Grouping	Students, Instuctors, or Others		
	Reliability		Equiva-
Measure	Computer-Based	Paper-and-Pencil	lence
Score	.53	.62	.50
Confidence	.94	.95	.76
Latency	.88	-	-

Note. Split-half reliability estimates were adjusted by employing the Spearman-Brown Prophecy Formula.

Table 2

Statistics Associated with Significant Discriminant Function, Standardized Discriminant-Function Coefficients, Pooled Within-Groups Correlations Between the Discriminant Function and Computer-Based and Paper-and-Pencil Measures, and Group Centroids for Above or Below F-14 or E-2C Mean Flight Hours

Discriminant Function

Eigen-value	Percent Variance	Canonical Correlation	Wilks Lambda	Chi Squared	d.f.	p
.44	82.43	.55	.64	31.38	15	.008

Measure	Discriminant Coefficient	Within-Group Correlation	Group	Centroid
CTS	.91	.51	Above F-14 Mean Hours	.10
CTC	.84	.57	Below F-14 Mean Hours	.39
CTL	-.24	-.45	Above E-2C Mean Hours	-1.35
PTS	-1.19	-.00	Below E-2C Mean Hours	-1.50
PTC	-.17	.36		

Table 3

Means and Standard Deviations for Groups Above or Below F-14 or E2-C Mean Flight Hours, Univariate F-Ratios, and Levels of Signifigance for Computer-Based and Paper-and-Pencil Measures

Measure		Group				F	p
		Above F-14 Flight Hours (n=26)	Below F-14 Flight Hours (n=37)	Above E-2C Flight Hours (n=5)	Below E-2C Flight Hours (n=7)		
CTS	$\bar{X}$	60.58	59.62	44.60	43.14	2.94	.039
	s	15.75	18.77	15.68	17.37		
CTC	$\bar{X}$	75.58	80.84	48.60	64.57	4.11	.010
	s	21.57	19.80	21.23	26.48		
CTL	$\bar{X}$	8.42	7.81	9.49	11.06	2.28	.087
	s	3.31	2.77	4.10	3.94		
PTS	$\bar{X}$	51.65	49.73	45.80	52.86	.19	.900
	s	18.26	20.38	11.86	13.91		
PTC	$\bar{X}$	72.23	76.70	53.00	69.71	2.14	.103
	s	23.02	18.10	16.55	20.94		

Table 4

Statistics Associated with Significant Discriminant Function, Standardized Discriminant-Function Coefficients, Pooled Within-Groups Correlations Between the Discriminant Function and Computer-Based and Paper-and-Pencil Measures, and Group Centroids for F-14 RIOs or Pilots and E-2C NFOs or Pilots

Discriminant Function						
Eigen-value	Percent Variance	Canonical Correlation	Wilks Lambda	Chi Squared	d.f.	p
.66	81.96	.63	.53	44.72	15	.000

Measure	Discriminant Coefficient	Within-Group Correlation	Group	Centroid
CTS	-.73	-.48	F-14 RIOs	-.32
CTC	-.32	-.52	F-14 Pilots	-.21
CTL	.57	.58	E-2C NFOs	.58
PTS	-1.15	-.05	E-2C Pilots	3.13
PTC	-.45	-.45		

Table 5

Means and Standard Deviations for Groups of F-14 RIOs or Pilots and E2-C NFOs or Pilots, Univariate F-Ratios, and Levels of Signifigance for Computer-Based and Paper-and-Pencil Measures

Measure		Group				F	p
		F-14 RIOs (n=37)	F-14 Pilots (n=26)	E-2C NFOs (n=8)	E-2C Pilots (n=4)		
CTS	$\bar{X}$	60.57	59.23	48.88	33.50	3.74	.015
	s	17.46	17.77	9.11	23.01		
CTC	$\bar{X}$	79.78	77.08	65.50	42.75	4.39	.007
	s	20.67	20.66	18.80	31.08		
CTL	$\bar{X}$	8.18	7.88	8.40	14.43	5.84	.001
	s	3.42	2.30	2.49	3.00		
PTS	$\bar{X}$	50.68	50.31	51.38	47.00	.05	.984
	s	19.87	19.11	11.78	16.79		
PTC	$\bar{X}$	76.54	72.46	72.38	43.50	3.42	.022
	s	21.72	18.11	11.44	21.63		

Table 6

Statistics Associated with Significant Discriminant Function, Standardized Discriminant-Function Coefficients, Pooled Within-Groups Correlations Between the Discriminant Function and Computer-Based and Paper-and-Pencil Measures, and Group Centroids for VF-124 Students and Instructors or Members of Other Operational Squadrons

Discriminant Function

Eigen-value	Percent Variance	Canonical Correlation	Wilks Lambda	Chi Squared	d.f.	p
1.43	97.69	.77	.40	64.40	10	.000

Measure	Discriminant Coefficient	Within-Group Correlation	Group	Centroid
CTS	.62	.86	Students	1.34
CTC	.50	.70		
CTL	.02	-.32	Instructors	.05
PTS	.24	.67		
PTC	-.45	-.45	Others	-1.20

Table 7

Means and Standard Deviations for Groups of VF-124 Students and Instructors or Members of Other Operational Squadrons,Univariate F-Ratios, and Levels of Signifigance for Computer-Based and Paper-and-Pencil Measures

Measure		Group			F	p
		Students (n=30)	Instructors (n=11)	Others (n=34)		
CTS	$\bar{X}$	72.33	57.36	44.26	38.30	.000
	s	13.30	16.30	11.03		
CTC	$\bar{X}$	91.10	78.91	60.29	25.06	.000
	s	11.83	16.22	21.52		
CTL	$\bar{X}$	7.30	7.50	9.73	5.63	.005
	s	2.80	2.50	3.41		
PTS	$\bar{X}$	63.97	48.27	39.18	23.09	.000
	s	13.81	18.33	14.00		
PTC	$\bar{X}$	85.03	75.36	61.44	14.37	.000
	s	16.99	14.61	18.99		

DISTRIBUTION LIST

Assistant for Manpower Personnel and Training Research and Development (OP-01B2)
Head, Training and Education Assessment (OP-11H)
Cognitive and Decision Science (OCNR-1142CS)
Technology Area Manager, Office of Naval Technology (Code-222)
Office of Naval Research, Detachment Pasadena
Technical Director, U.S. ARI, Behavioral and Social Sciences, Alexandria, VA (PERI-ZT)
Superintendent, Naval Postgraduate School
Director of Research, U.S. Naval Academy
Institute for Defense Analyses, Science and Technology Division
Center for Naval Analyses, Acquisitions Unit
Department of Psychological Sciences, Purdue University
Defense Technical Information Center (DTIC) (2)